Discussion

Corpus (francophones) annotés en Entités Nommées: bilan et perspectives

JRC
EUROPEAN COMMISSION

CORPUS ANNOTÉS EN ENTITÉS NOMMÉES

Pourquoi ?

- **Mise au point d'outils d'annotation automatique**

  i.e. description précise des phénomènes à traiter pour :

  - guider le travail d'écriture de règles (systèmes symboliques)
  - apprendre automatiquement des règles (machine learning)

- **Evaluation**

  disposer d'une annotation de référence

→ ressources indispensables

JRC
EUROPEAN COMMISSION

CORPUS ANNOTÉS EN ENTITÉS NOMMÉES

Comment ?

- **Annotation manuelle**
  - définition de la tâche
  - rédaction de guides d'annotation précis
  - annotation (annotateurs, interface, etc.)
  - évaluation: accord inter annotateurs

- **Annotation automatique**
  - utilisation de Wikipédia

- **Crowdsourcing**
  - microworking de type "Amazon Mechanical Turk"

→ **Questions:**

Quelle qualité? Comment valider?
Quelle quantité?
Quel coût?

JRC
EUROPEAN COMMISSION

CORPUS ANNOTÉS EN ENTITÉS NOMMÉES

Quoi?

- **Les “classiques”**
  MUC, CONNL, ACE, IREX, HAREM, Evalita
- **Différents domaines**
  BioCreAtIvE, JNLPBA
- **Différentes tâches**
  SemEval 2007 (métonymie), WePS

JRC
EUROPEAN COMMISSION

CORPUS ANNOTÉS EN ENTITÉS NOMMÉES

Nouveaux besoins?
Nouveaux challenges?

- **Nouveaux types de textes**
  - textes écrits
  - RAP
  - OCR

  → Problème de qualité, données souvent bruitées

- **Nouvelles annotations**
  - Catégories plus étendues et plus variées
  - Caractérisation plus précise → annotation plus fine
  - Annotation “référentielle”

  → Tâche d’annotation encore plus complexe

- **Toujours plus de langues**

Constitution de corpus annotés multilingues par projection cross-linguistique

- **Principe**
  Etant donné un set de corpus parallèles multilingues, annotation des EN pour une langue puis projection sur les autres langues

- **Méthode**
  - annotation automatique de l'anglais
  - traduction des EN "source" (SMT)
  - projection : processus itératif utilisant trois méthodes (strict matching, consonant signature, distance similarity)

- **Expériences**
  - En, Fr, Es, De, Cz / En, Ru (2000 entités)
  - expansion:
    - + italien et hongrois
    - + de données
    - annotation source : système de vote combinant 3 systèmes